An Urban Geography of Mobile Application Usage: Connecting Demand Dynamics and Urban Fabrics

Sachit Mishra*†, Diego Madariaga*, Cezary Ziemlicki‡, Diala Naboulsi§, Marco Fiore*
*IMDEA Networks Institute, Spain, †Universidad Carlos III de Madrid, Spain,
‡SENSE / Orange Innovation, France, §École de Technologie Supérieure, Canada
{sachit.mishra, diego.madariaga, marco.fiore}@imdea.org, cezary.ziemlicki@orange.com, diala.naboulsi@etsmtl.ca

***Abstract*—The surge in usage of mobile applications generates a massive volume of traffic data exhibiting unique dynamics that are hard to unravel. In this work, we leverage factor analysis to pin down recurrent patterns of mobile traffic over the three dimensions of space, time and services in multi-city measurements of unprecedented resolution. We link the revealed structures of real-world mobile demands to urban fabrics, *i.e.*, the combination of infrastructures and social characteristics that determine the functionality of an urban territory, hence establishing connections between specific city landscapes and the mobile application consumption they create. Our study provides new understanding about the diversity of mobile service dynamics in metropolitan areas, including insights on how economic status drives adoption of specific applications, how residential versus commercial areas create a dichotomy in applications usage, how private and public transports drive surges in the prevalence of different sets of applications, or how nightlife or university studies stimulate the utilization of specific classes of services.**

I. Introduction

The rapid development of mobile communication technologies in the last two decades has created unprecedented changes in our society. People rely today on mobile applications to manage many aspects of their lives, including getting informed and forming opinions, organizing their daily personal and professional activities, moving with private and public transport, spending their free time or even dating. The entanglement of mobile applications with real-world life is especially strong in dense urban areas, where individuals are known to make a substantially higher use of such technologies [1]. In fact, the penetration of mobile communications in cities is so high that it has become a recognized tool to investigate socio-economic urban phenomena like mobility [2]–[5], poverty [6], [7], inequality [8], or disease transmission [9]–[11].

Prompted by these considerations, in this paper we investigate the existence of connections between the spatiotemporal consumption patterns of mobile applications and the underlying *urban fabrics*, *i.e.*, the entangled blend of physical infrastructure, human activities and socio-economic characteristics that combine to determine the functionality of an urban territory. Our aim is therefore that of (i) identifying recurrent utilization schemes of one or more mobile applications over space or time and (ii) linking such schemes to the urban fabric that explains their emergence. This approach lets us establish causal relationships between the city landscape and the complex mobile service demands it generates, ultimately revealing why people cling to specific mobile applications at different times and locations. The result has clear social implications [12] as well as important applications to networking, where it can, *e.g.*, drive a better urban-fabrics-informed planning of networks supporting application-specific operations such as slicing [13].

Several studies have examined the link between mobile network traffic and urban environments, but often with different goals or datasets. For example, analyses of 3G [14], 4G [15], and 5G [16], [17] networks include basic geographical insights but lack the in-depth connections we pursue. Similarly, prior works on traffic patterns of app categories (e.g., video streaming, messaging) [18]–[24] or of specific software like WhatsApp or Facebook [25]–[27], focus on service demand locality without deeper urban correlation analysis. Closer to our goal, a significant body of works has focused on land use detection via mobile data [28]–[33], thus relating network traffic to the underlying topography. Yet these analyses typically employ mobile call and texting records that are especially suitable to tell apart land use categories (*e.g.*, residential, agricultural, commercial, retail, industrial); instead, our target is the traffic generated by individual applications and we aim at associating it to more general urban fabrics that are characterized not only by land usages but also by combinations of urban infrastructures and socio-economic status.

Other studies that have explored the interface of application demands and geographical locations have done so with very different objective than ours. For instance, previous works have validated urban planning theories [34], characterized indoor cellular traffic [35], profiled users based on their spatiotemporal mobile service demands [36], [37], or predicted application usage based on user location [38], [39]. However, none has focused on the causality relationships between the urban environment and the mobile application consumption.

There exist also a few investigations that have explicitly coupled geographical features with the mobile demands they entail. Yet, they have considered aggregate data traffic over all services that cannot explain the spatiotemporal behavior of each application nor its connection with urban fabrics [40], [41]; or, they have carried out application-level analyses at countrywide scales and with low spatial resolutions (*e.g.*, at the granularity of large administrative units covering whole cities) that do not allow inspecting urban dynamics [42], [43].

Ultimately, we still lack a clear understanding of the exact connections between mobile service usages and the urban features that determine them. We contribute to closing such a knowledge gap with the following main contributions.

- We tailor exploratory factor analysis to the problem of identifying complex recurrent patterns in the demands for mobile services, which lets us extract hidden behaviors that affect either a single service or a whole set and that occur over space, time or both dimensions at once.

- We apply such a factor analysis to a state-of-the-art measurement dataset of mobile network traffic, capturing the usage of tens of applications at a high spatial resolution of 100×100 m^2 and encompassing multiple urban areas in a major European country.
- We identify tens of application-specific spatiotemporal dynamics that occur across all studied urban areas and link those to precise urban fabrics, unveiling the inherent and previously unknown connections between the city territory and the mobile traffic generated by its inhabitants.

II. Data measurement and processing

Our study builds upon network traffic measurements collected in the production network of Orange, a major global operator with a leading market position in France. The data was collected by jointly monitoring the 4G Radio Access Network (RAN) and Core Network (CN) infrastructure that provides coverage to the main urban areas of France during 10 consecutive weeks. The resulting traffic measurements concern ten major French cities, *i.e.*, Paris, Lyon, Marseille, Toulouse, Bordeaux, Strasbourg, Nantes, Nice, Le Mans, and Dijon.

A. Network monitoring platform

The traffic measurements were performed by the network operator using passive measurement probes tapping at the Gi, SGi and Gn interfaces connecting the Gateway GPRS Support Nodes (GGSNs) and Packet Data Network Gateways (PGWs) of the Long Term Evolution (LTE) Evolved Packet Core (EPC) network to external public data networks (PDNs). This monitoring strategy allows capturing the 4G traffic traversing the mobile network across the whole country. The probes run dedicated proprietary classifiers that allow associating individual TCP and UDP flows to the mobile applications that generate them, for purposes that include network monitoring, traffic engineering, and research activities. The dataset used for our study contains information about 68 mobile services.

To geographically localize the measured traffic we associate it to the serving base stations. Specifically, we resort to Network Signaling Data (NSD) captured by probes monitoring the S1 interface connecting base stations to the Mobility Management Entity (MME). NSD events allow associating each traffic flow to the exact sequence of its servicing antennas, which lets us allocate the correct fraction of the total volume of data traffic in the flow to each serving base station. The flows association to base stations are updated at every 15 minutes, hence the dataset describes mobile applications traffic records at each base station with that same temporal granularity.

B. Traffic interpolation to statistical zones

In order to ease analyses throughout the study, we perform an interpolation of the traffic collected at each base station onto a zoning defined by the French National Institute of Statistics and Economic Studies (INSEE). The spatial interpolation is illustrated in Figure 1, and involves the following steps.

(A) Definition of the empirical coverage matrix of each base station at a high spatial resolution. Coverage is encoded as the probability $P(\ell|i)$ that a user served by base station i is located at a given *tile* ℓ of 100×100 m^2. The information is provided to us by the network operator.

(B) Extraction of the traffic time series at each base station. For each base station i, we compute the 10-week times series $\mathcal{T}_a^i(t)$ of each mobile services a with a 15-minute temporal granularity of time t.

(C) Geographical mapping of the temporal traffic. For each base station i we multiply the traffic time series $\mathcal{T}_a^i(t)$ with the coverage matrix $P(\ell|i)$, obtaining a spatiotemporal representation $\mathcal{M}_a^i(\ell,t)$ of the demand at i.

(D) Computation of the aggregate spatial traffic. The spatiotemporal maps of traffic from all base stations at a given time t are summed to obtain the overall service-level traffic maps $\mathcal{M}_a(\ell,t)$ for each application a.

(E) Interpolation to statistical zoning. We retrieve the *IRIS zoning* produced by INSEE, which tessellates the urban territory of each target city into statistical zones based on geographical and demographic criteria; IRIS zones have homogeneous surfaces and encompass populations of at most $2,000$ local inhabitants. We then assign to each IRIS zone the demands $\mathcal{M}_a(\ell,t)$ associated to all tiles ℓ it covers. The traffic of tiles overlaying multiple IRIS zones is distributed proportionally to the fraction of tile covered by each overlapping IRIS zone.

The result of this processing are service-level time series $\mathcal{D}_a^z(t)$ of the traffic demand for each application a at each of 5,097 IRIS zones z covering the ten major cities under consideration.

C. Ethics considerations

The measurements run to collect the data used for our study were performed by the operator for network management and research purposes, and temporarily stored within a secure platform at their own premises. The raw data processing was carried out in the same platform by personnel of the network operator, in full compliance with Article 89 of the General Data Protection Regulation (GDPR) [44] of the European Commission. The data collection and processing were approved by the Data Protection Officer (DPO) of the operator, and authorized by the relevant national privacy-protection agency. The researchers involved in our study only had access to the aggregate data $\mathcal{T}_a^i(t)$ and $P(\ell|i)$, whose spatiotemporal resolution ensures that no data subject can be re-identified from the data, which in fact does not qualify as personal data in the GDPR acceptation.

III. Refinement of Mobile Services

Each IRIS zone (hereinafter also simply referred as *zone*) in the dataset is associated with the traffic demands of 68 individual mobile services over 10 weeks at a temporal resolution of 15 minutes. However, given the high diversity of popularity and penetration of mobile applications, not all of them are statistically relevant to an analysis at a IRIS zoning resolution; indeed, many services have too few users

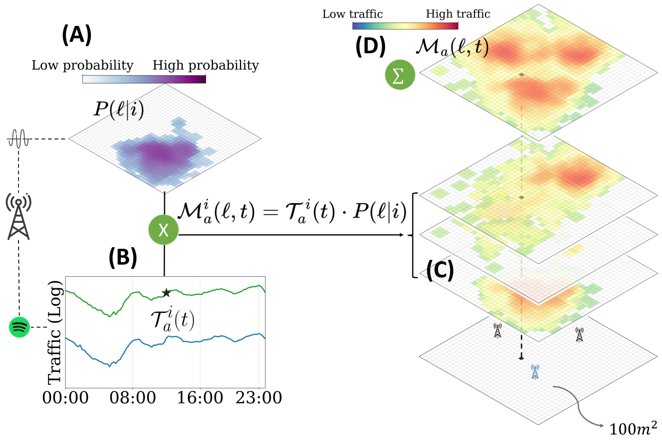
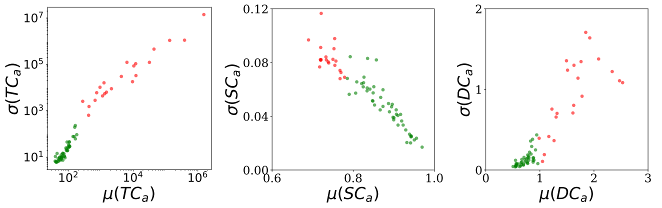


Fig. 1: Methodology for computing the traffic maps; (a) Coverage matrix for an base station i, (b) Traffic time series for Spotify, (c) Multiplication of the coverage matrix with the traffic time series, (d) Summation of all traffic maps.



(a) Time (b) Space (c) Dispersion

Fig. 2: Filtering of applications with inconsistent dynamics over (a) time, (b) space and (c) dispersion. Accepted services in each plot are colored in green and identified via DBSCAN.

or generate negligible and essentially random demands within each zone over 15 minutes. To tell apart the services that generate meaningful patterns from those that are too noisy to yield usable information, we change the resolution of time series $\mathcal{D}_a^z(t)$ to 1 hour by aggregating consecutive 15-minute time slots. Then, we define the *median week* time series of application a in zone z as

$$\tilde{\mathcal{D}}_a^z(\tau) = \mathsf{med}\left(\mathcal{D}_a^z(t) \ : \ t \,\mathsf{mod}\, T = \tau\right), \tag{1}$$

where the med and mod operators denote the median of a set and the modulo operation, respectively, whereas T=168 is the number of hours in one week. The expression in (1) returns the median traffic volume for service a in zone z at each hour τ of a complete week.

As such, $\tilde{\mathcal{D}}_a^z(\tau)$ already de-noises the time series via the median calculation. We next explore if such a process is sufficient to make each service a show consistent and usable demand dynamics across zones. More precisely, we define metrics to test the quality of the service-level median week time series in terms of their time, space, and data dispersion, as defined next. The ultimate goal is identifying the subset of reliable applications A^* to be adopted in the rest of our study.

A. Temporal consistency

In order to assess the temporal regularity of the demands, we adapt the Total Variation (TV) metric [45] to measure the changes between consecutive values in the $\tilde{\mathcal{D}}_a^z(\tau)$ time series for each service a and zone z. The original TV definition is

$$\mathrm{TV}(\tilde{\mathcal{D}}_a^z(\tau)) = \sum_{\tau=1}^{T} \left|\tilde{\mathcal{D}}_a^z(\tau+1) - \tilde{\mathcal{D}}_a^z(\tau)\right|, \tag{2}$$

which effectively measures the temporal smoothness of $\tilde{\mathcal{D}}_a^z(\tau)$ but cannot be used to compare the smoothness among time series of differences in magnitudes that often characterize $\tilde{\mathcal{D}}_a^z(\tau)$. We introduce the Total Percentage Variation (TPV) metric, which quantifies consecutive relative differences instead of absolute differences, as follows

$$\mathrm{TPV}(\tilde{\mathcal{D}}_a^z(\tau)) = \sum_{\tau=1}^{|T|-1} \frac{\left|\tilde{\mathcal{D}}_a^z(\tau+1) - \tilde{\mathcal{D}}_a^z(\tau)\right|}{\tilde{\mathcal{D}}_a^z(\tau)}. \tag{3}$$

The expression allows studying the time consistency of a service a by composing the set $TC_a = \{\mathrm{TPV}(\tilde{\mathcal{D}}_a^z(\tau)) \mid z \in Z\}$ over the set Z of all zones and then computing its mean $\mu(TC_a)$ and standard deviation $\sigma(TC_a)$. Intuitively, time consistent mobile services have low $\mu(TC_a)$ and $\sigma(TC_a)$ values, indicating smooth demands opposed to bursty and random traffic. Figure 2a shows the distribution of pairs $(\mu(TC_a)$,$\sigma(TC_a))$: as expected, applications with higher average TC_a also tend to display an increasing deviation of the metric, pinpointing abrupt and high changes in the traffic that translate into noise for our study. In order to separate the subset of time-consistent applications, we apply the density-based clustering algorithm (DBSCAN) [46] on the mean and deviation pairs of all applications, which returns a cluster of 49 applications with time regularity (green in the figure).

B. Spatial consistency

To determine the spatial regularity of mobile service demands, we analyze the level of correlation of the service-level traffic demands recorded within a specific zone and the adjacent zones. Formally, let N_z be the neighborhood set of z that contains the K zones that are the geographically closest to z. We define the Local Correlation (LC) metric as

$$\mathrm{LC}(\tilde{\mathcal{D}}_a^z(\tau)) = \frac{1}{K} \sum_{\tilde{k}\in N_z} \rho\big(\tilde{\mathcal{D}}_a^z(\tau), \tilde{\mathcal{D}}_a^k(\tau)\big), \tag{4}$$

where ρ is the Pearson correlation. The LC values are in $[-1, 1]$, with values closer to 1 suggesting a strong positive correlation among neighbors. The space consistency of a service a is described by the set $SC_a = \{\mathrm{LC}(\tilde{\mathcal{D}}_a^z(\tau)) \mid z \in Z\}$. Figure 2b shows the distribution of pairs $(\mu(SC_a)$, $\sigma(SC_a))$, where spatially consistent services are related to high $\mu(SC_a)$ but low $\sigma(SC_a)$, *i.e.*, strong local correlation across all zones. Again, we observe that less consistent applications that are characterized by lower mean correlations LC_a tend to also have larger standard deviations, which detects services that have fairly random behaviors in contiguous zone and may hamper spatial patterns in our study. Applying DBSCAN to the LC_a statistics returns a cluster of 48 mobile services that yield space consistency (green in the figure).

C. Data dispersion consistency

We refer to data dispersion as the variability in $\tilde{\mathcal{D}}_a^z(\tau)$ in relation to its mean value. To quantify the level of dispersion of a service-level traffic time series we use the Coefficient of Variation (CV) [47], which is defined as the ratio of the standard deviation to the mean, formally

$$\mathrm{CV}(\tilde{\mathcal{D}}_a^z(\tau)) = \frac{\sigma\big(\tilde{\mathcal{D}}_a^z(\tau)\big)}{\mu\big(\tilde{\mathcal{D}}_a^z(\tau)\big)}. \tag{5}$$

The CV is a dimensionless quantity, and it can be employed to compare the dispersion of data between time series free from scale effects. In addition, CV values above 1 indicate high variability, while values below 1 indicate low data variability, *i.e.*, high data dispersion consistency [47].

We describe the data dispersion consistency of a mobile application a by the set $DC_a = \{\mathrm{CV}(\tilde{\mathcal{D}}_a^z(\tau)) \mid z \in Z\}$, with mean $\mu(DC_a)$, and standard deviation $\sigma(DC_a)$. As in the previous approaches, Figure 2c exhibits the distribution of pairs $(\mu(DC_a),\ \sigma(DC_a))$, resulting in a unique cluster identified by DBSCAN, grouping a subset of 46 mobile services with an adequate level of data dispersion consistency.

D. Application selection

Intersecting the clusters of consistent applications returned by each of the previously described criteria, we obtain a subset A^* of 30 mobile services that fulfill our requirement of reliability in time, space, and data dispersion. The subset A^* includes, *e.g.*, Instagram, YouTube, Facebook, Netflix, Twitter/X, SnapChat, or Spotify, just to name a few.

IV. Factor analysis of service-level demands

In order to explore the underlying structure of mobile application usage, we rely on an Exploratory Factor Analysis (EFA) approach [48]. We cast EFA so as to identify the *common factors* that explain specific spatial and temporal patterns in the mobile usage dynamics at service level, as detailed next.

A. EFA operation

In its fundamental definition, EFA assumes that the structure of a large set of observable variables can be modeled via a linear combination of unobservable common factors and error terms. Considering a set of N samples described by p variables with a set of k common factors, the EFA model can be expressed, in matrix notation, as

$$\mathbf{X} = \mathbf{M} + \mathbf{LF} + \varepsilon, \tag{6}$$

where

- $\mathbf{X} = (x_{ij})_{p\times N}$ is the data matrix,
- $\mathbf{M} = (m_{ij})_{p\times N}$ is the mean matrix, $m_{ij} = \frac{1}{N}\sum_{l=1}^{N} x_{il}$,
- $\mathbf{L} = (l_{ij})_{p\times k}$ is the factor loading matrix that describes the relationship between variables and common factors,
- $\mathbf{F} = (f_{ij})_{k\times N}$ is the factor score coefficient matrix that describes the sample placement on the factor distribution,
- ε is the error term matrix that yields the residual behavior of each sample not explained by the common factors.

A simplified EFA model can be derived from (6) as follows

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\intercal + \mathbf{\Psi}, \tag{7}$$

where $\mathbf{\Sigma} := \mathrm{Cov}(\mathbf{X} - \mathbf{M})$ is the covariance matrix from the data, and $\mathbf{\Psi} := \mathrm{Cov}(\varepsilon)$ is the covariance of the error term.

In our case, EFA is employed to summarize the traffic patterns of mobile applications observed at a large number of locations. Thus, the input for EFA is a set of zones (*i.e.*, samples), each characterized by different traffic features (*i.e.*, variables) that describe the consumption of mobile services. The objective of (7) becomes then identifying a limited set of common factors that explain the largest portion possible of the total variance in the traffic demand data across zones.

We proceed by engineering the traffic features fed to EFA as well as the implementation of the the different steps required to fit (7) to the measurement data and to interpret the results.

B. Service-level demand signature

The raw data matrix $\mathbf{X}$ is composed by p=30×4×24×7×10 variables for $N = \|Z\|$ zones, since in each zone we have the traffic generated at every 15 minutes for 10 weeks by 30 different services upon the filtering in Section III. This volume of information is too large to be ingested as is for downstream analysis and would render (i) the resolution of the EFA representation computationally unfeasible and (ii) the interpretation of the result extremely involved since the resulting loading factor matrix $\mathbf{L}$ has also a dimension p.

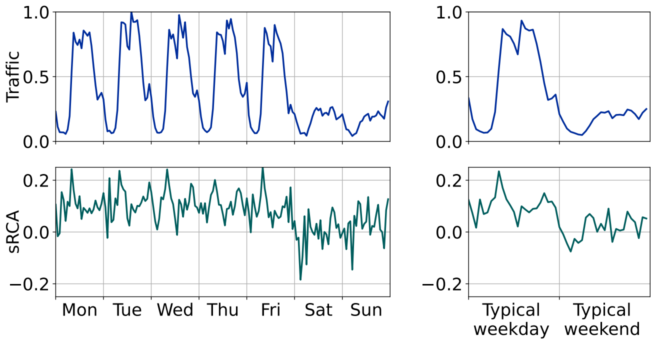
Instead, we propose a *demand signature* that compresses the voluminous data above into a much more compact format that still retains the key spatiotemporal dynamics of the traffic. The signature can thus be effectively used as the input feature, cutting complexity and preserving the quality of the insights.

Our demand signature definition is twofold, as follows.

- It captures the individual temporal dynamics of the traffic generated by each application a in the target zone z as the demand $\mathcal{D}_a^z(t)$ already introduced in Section II.
- It represents the relative prevalence of the traffic generated by each application a in the target zone z with respect to all other services as the revealed comparative advantage (RCA) index [49]. The RCA is defined as

$$\mathcal{R}_a^z(t) = \frac{\mathcal{D}_a^z(t)/\mathcal{D}_A^z(t)}{\mathcal{D}_a^Z(t)/\mathcal{D}_A^Z(t)}, \tag{8}$$

where $\mathcal{D}_A^z(t)$, $\mathcal{D}_a^Z(t)$ and $\mathcal{D}_A^Z(t)$ denote the aggregate traffic demands at time t of (i) all services at z, (ii) application a over all zones, and (iii) all services over all zones (*i.e.*, the total traffic recorded in our data at t), respectively. The RCA thus quantifies how relative consumption of an application a within a zone z at t (numerator) compares against the typical consumption of that service in all zones at the same time (denominator). In fact, the RCA defined in (8) takes values in $[0, \infty]$. To avoid issues related to the unbounded nature of such, we consider a variant of the original RCA definition, *i.e.*,



(a) Median week signature (b) Condensed signature

Fig. 3: Demand signature of Apple Music within a zone z.

the symmetric revealed comparative advantage (SRCA), formall defined as follows

$$\mathcal{S}_a^z(t) = \frac{\mathcal{R}_a^z(t) - 1}{\mathcal{R}_a^z(t) + 1}. \tag{9}$$

The SRCA values are in $[-1, 1]$, with values lower than 0 indicating under-consumption and values above 0 indicating over-consumption with respect to the typical activity over the whole monitored territory at time t.

While $\mathcal{D}_a^z(t)$ and $\mathcal{S}_a^z(t)$ offer complementary perspectives on service-level demands, they do not solve and actually worsen the problem of the variable set size, which is now multiplied by two. We compress $\mathcal{D}_a^z(t)$ and $\mathcal{S}_a^z(t)$ in two steps.

First, similarly to Section III, we compute median weeks of both time series for all applications and zones. For the traffic $\mathcal{D}_a^z(t)$, we compute $\tilde{\mathcal{D}}_a^z(t)$ as per equation (1); in addition, since we are interested in analyzing consumption patterns instead of actual traffic volumes that can vary dramatically across zones, we min-max normalize all $\tilde{\mathcal{D}}_a^z(t)$ time series. An example of resulting traffic median week time series is in the top plot of Figure 3a for one representative zone z and application a. For the SRCA, we instead compute

$$\tilde{\mathcal{S}}_a^z(\tau) = \mathtt{med}\left(\mathcal{S}_a^z(t) \; : \; t \bmod T = \tau\right), \tag{10}$$

an example being in the bottom plot of Figure 3a.

Second, we further reduce the dimensionality of the signature, which is still at p=30×168×2, *i.e.*, above 10,000. Based on well-known regularities in mobile traffic [31], we take advantage of the fact that the most distinguishing dynamics tell apart working days (Monday to Friday in France) and weekends (Saturday and Sunday in France) to generate a more concise representation of $\tilde{\mathcal{D}}_a^z(\tau)$ and $\tilde{\mathcal{S}}_a^z(\tau)$. The two time series are reduced to two periods of 24 hours each, modeling working days and weekends respectively, by averaging each hour of the day over the corresponding days in $\tilde{\mathcal{D}}_a^z(\tau)$ and $\tilde{\mathcal{S}}_a^z(\tau)$. The result is illustrated in Figure 3b for the same zone z and application a of Figure 3a.

C. Suitability of the data for EFA

The resulting demand signature has a size p=30×48×2, *i.e.*, a sample-to-variable ratio N/p of roughly 2:1 considering that we have N=$\|Z\|$=5,097 zones in our dataset. This complies with rules of thumb adopted by previous works based on EFA that recommend N/p ratios higher than 1. Yet, there is no evidence of a minimum ratio needed to achieve satisfactory factor recovery [50]; in fact, the optimal N/p value highly depends on the domain and data over which EFA is applied [51].

In order to formally verify the suitability of the demand signature for EFA, we therefore rely on the widely used Kaiser-Meyer-Olkin (KMO) test [52]. The KMO test is based on the correlation between the observed variables and indicates how well each variable can be explained by the other variables. KMO values range between 0 and 1, with values closer to 1 suggesting that EFA should produce reliable factors. We apply the KMO test to our data composed of $N = 5,097$ samples and $p = 2,880$ variables, resulting in a value of 0.988 and indicating that the dataset is highly suitable for EFA.

It is also reasonable to question why we prefer EFA over the more commonly used Principal Component Analysis (PCA) method, which addresses the similar problem of reducing the number of variables to fewer items. In fact, EFA and PCA greatly differ in their goals. PCA aims at reducing the number of variables into components that explain as much of the variance as possible. On the other hand, EFA identifies a latent structure of common factors that only explain the common variance of the data and not the unique variance. Ultimately, PCA is a tool for dimensionality reduction while EFA shall be used to identify latent correlations [53]. As the latter is our objective, EFA is the appropriate method whereas choosing PCA would incur into less informative results [54].

D. Factor extraction

There are multiple methods to fit an EFA model, among which we select the Maximum Likelihood estimator (MLE) [55] for our data. The main goal of MLE is to enhance the interpretability of the common factors while providing an adequate fit to the data as per (7). The MLE approach is popular for extracting common factors, and is especially apt in use cases with a large number of samples like ours.

The MLE method produces parameter estimates, *i.e.*, $\mathbf{L}$ and $\mathbf{\Psi}$ in (7)) that minimize the discrepancy between $\mathbf{\Sigma}$, the covariance matrix obtained from the data matrix, and $\hat{\mathbf{\Sigma}}$, the covariance matrix implied by the hypothesized model. Thus, the MLE factor extraction is obtained by iteratively minimizing the following discrepancy function:

$$f_{\mathrm{MLE}} = \mathrm{tr}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}) - \log|\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}| - p, \tag{11}$$

where tr is the trace function that sums the elements on the diagonal of a square matrix.

E. Factor retention

A critical decision in an EFA is selecting the optimal number k of common factors. Intuitively, a higher k improves the fit of the data, but it may also lead to overfitting of the model and a more complex interpretation of the resulting factors. Several statistical methods have been proposed in the literature to determine how many factors to include in an EFA.

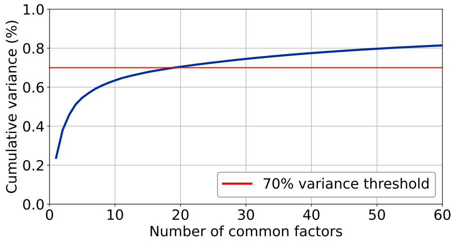


Fig. 4: Percentage of explained variance versus k.

Nevertheless, most of these criteria have not been thoroughly tested on datasets beyond social sciences domains.

We thus experimented with some of the most widespread methodologies for our study, including Parallel Analysis, the Empirical Kaiser Criterion, and the Minimum Average Partial method. However, they all suggested retaining more than 100 factors, a clear indicator of so-called *overfactoring*. Finally, we employ one of the fundamental methods for factor retention, *i.e.*, the Cumulative Percentage of Variance (CPV). Thus, we select the number of factors based on how much variability can be explained as we include more factors. Figure 4 shows the cumulative variance explained in our data as a function of the number of factors retained. We set the desired amount of CPV to 70%, resulting in the retention of 20 factors. The selected value aligns with the literature, as the proposed thresholds for the CPV typically range between 60% and 95%, depending on the field of study and target data [56].

F. Factor rotation

One of the properties of the EFA model is its rotation invariance: the factor loading matrix L can be rotated within the variable space, preserving the same fit quality. This property can be exploited to enhance the factors by reorganizing and simplifying their structure, which translates into a more interpretable rotated factor loading matrix $\mathbf{L}^* = (l^*_{ij})$. In our analysis, we employ the Varimax approach [57], one of the most used rotation methods. Varimax is a type of orthogonal rotation that assumes factors to be uncorrelated and minimizes the number of variables with high loadings on each factor. More precisely, Varimax computes scaled loadings $\tilde{l^*_{ij}} = l^*_{ij}/h_i$, where $h_i = \sum_{j=1}^{k} l^2_{ij}$. Then, the procedure finds the factor rotation that maximizes the following function:

$$V = \frac{1}{p}\sum_{j=1}^{k}\left\{\sum_{i=1}^{p}\left(\tilde{l^*_{ij}}\right)^4 - \frac{1}{p}\left(\sum_{i=1}^{p}\left(\tilde{l^*_{ij}}\right)^4\right)^2\right\} \qquad (12)$$

As a result, the differences between the loading factors are maximized in the rotated factor loading matrix.

V. Mobile application usage in urban France

Applying EFA to the service-level demands via the adaptations reported in Section IV yields 20 common factors across the 5,097 IRIS zones that compose the ten cities observed. Table I presents a summary for 8 of such 20 common factors and associates each to a distinctive urban fabric. For each factor we also report the applications whose demand dynamics are driving the factor, the temporal periods affected by the factor and a short description. Collectively, the factors we study account for 55.2% of the total variance in the data. The remaining 12 factors are associated with a minor fraction of the variance, and focus on less prominent urban fabrics, such as suburban areas or large parks in the city outskirts, hence we do not discuss them in detail due to space limitations.

Factor	Urban fabric	Applications	Time
1	Income level	All	All
2	Residential	All	8 am – 6 pm
4	Budget nightlife	All	10 pm – 6 am
6	University students	Instagram, Twitter/X	All
8	Main roadways	Waze, Siri	All
10	Train commuting	All	7-8 am, 6-7 pm
14	Airports	All	All
15	App-driven locations	Pokemon Go	All

TABLE I: Summary of 8 prominent common factors identified by EFA in the traffic generated by 30 mobile applications in 5,097 IRIS zones of 10 French cities.

We next delve into the analysis of the selected factors above. To this end, we leverage three distinct representations that can be extrapolated from the EFA model in (7) as follows.

- *Sum of the squared loadings* over each application a for the target factor k. Formally, let us define as $\boldsymbol{a}$ the subset of values in the demand signature in Section IV-B that refer to application a only; this is a set of 96 variables out of the complete p=2880 variables that compose the signature. Then the metric is $L^2_a(k) = \sum_{i\in\boldsymbol{a}} l^2_{ik}$ and is useful to understand which services have higher squared loadings and are therefore most concerned with the factor. An example of this metric is in Figure 5 for factor 1.
- *Loadings* separated by application for the target factor k. We display the actual loadings l_{ik} such that $i \in \boldsymbol{a}$, which as said above correspond to 96 variables, as a heatmap. These loadings indicate how the factor is linked to the typical hourly traffic demand and SRCA of application a in weekend and weekdays, as per our signature definition. Higher loadings towards 1 indicate that service a is highly affected by factor k at the specific hour, either in terms of its traffic dynamics $\tilde{\mathcal{D}}^z_a(\tau)$ or relative traffic share $\tilde{\mathcal{S}}^z_a(\tau)$. Figure 6 shows sample heatmaps for factor 1, where the top two rows refer to the SRCA of a in weekends (first row) and working days (second row) while the bottom two rows are associated to the normalized demand of a in weekends (third row) and working days (fourth row).
- *Scores* separated by application for the target factor k. We plot as a geographical map the scores f_{kj}, for all samples, *i.e.*, IRIS zones, $j \in Z$. Scores indicate how prevalent the factor k is in each zone, and allows us to clearly visualize areas of the observed cities that show high scores hence are dominated by the applications and traffic dynamics that characterize k. An example of scores map is in Figure 8a for factor 1 and zones in Paris.

A. Factor 1, or the digital divide of wealth

Factor 1 describes a common pattern across most of the monitored applications, as illustrated by the sum of squared loadings in Figure 5. Mail services (*e.g.*, Apple, Gmail) and

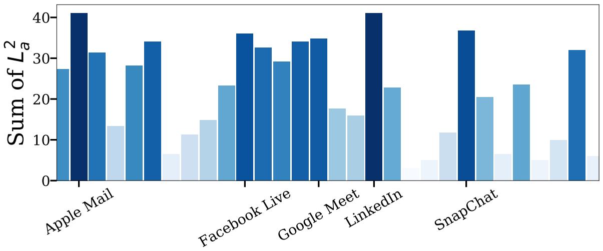


Fig. 5: Factor 1 sum of squared loadings for all applications.

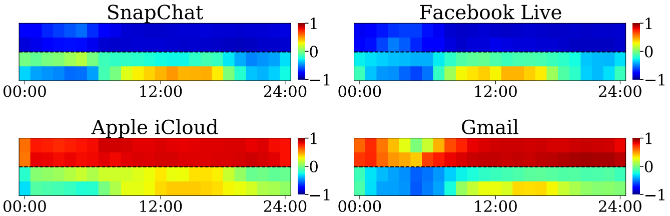


Fig. 6: Factor 1 loadings for selected applications. In each plot, SRCA (resp., traffic) loadings are top (resp., bottom) two rows.

LinkedIn yield especially high L_a^2 values, but the same holds for entertainment services like social media (*e.g.*, Facebook) or messaging (*e.g.*, SnapChat). Looking at the detailed loadings of selected applications with especially high L_a^2, in Figure 6 reveals that such applications are in fact affected by factor 1 in diametrically opposite ways. The factor is characterized by high loadings in the SRCA (red in the figure) of productivity-oriented applications like mail or cloud services, which thus experience much higher relative consumption than other services; it is instead associated with low SRCA loadings close to -1 (blue in the figure) for messaging or social media that are thus used much less than customary. The usage patterns above are consistent over time, as the SRCA loadings do not show temporal variance. The loadings for the normalized traffic dynamics, in the bottom two rows of each plot, more muddled in the case of factor 1.

Figure 8a shows the factor 1 scores across Paris, dividing the city into two distinct regions. When compared with the income distribution in Figure 8b, the similarity is evident, with a high Pearson correlation of 0.7 between the scores and INSEE's average income data per IRIS zone. Similar strong correlations are observed in other cities as well.

Takeaway. The geography of mobile service usage is predominantly is driven by wealth. Economic imbalance produces the single most important factor in our analysis, which explains alone over 20% of the total mobile traffic variance. High-income areas are characterized by a stable higher usage of a specific set of applications oriented at productivity, whereas low-income areas are characterized by a higher usage of social media and messaging services. The result reinforces works in the social sciences literature suggesting that notable consumption of social media is associated to lower education and income [58], and unveils the huge magnitude of the effect.

B. Factor 2, or the work-life dynamics of mobile traffic

Factor 2 affects all applications, with uniformly high L_a^2 values similarly to what happens in factor 1. The loadings however illustrate a completely different effect than factor 1. Figure 7 depicts the loadings of six applications, however we observed the same identical pattern in all monitored services. In this case, the RCA loadings are all around 0, implying that the factor is not characterized by a higher or lower relative usage of specific applications. Instead, the major dynamics occur in the loadings of the normalized traffic, in the bottom two rows of each plot. The factor is clearly characterized by a substantial drop of demands for all services (hence, lower total mobile traffic) during the working hours on Monday through Friday (as shown by the fourth row in each heatmap). The total network activity is instead especially high in the evening and during weekends.

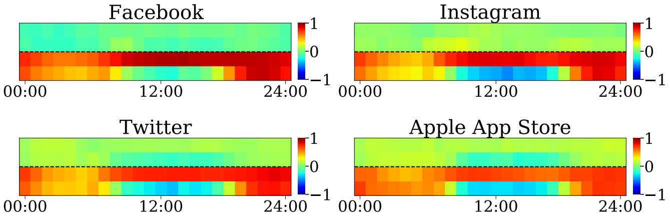


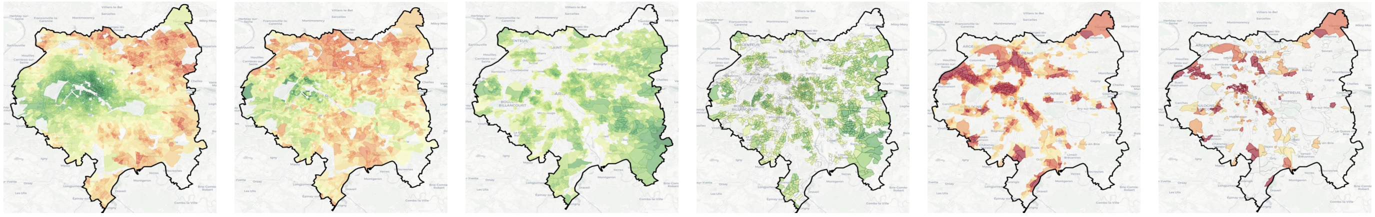
Fig. 7: Factor 2 loadings for selected applications. In each plot, SRCA (resp., traffic) loadings are top (resp., bottom) two rows.

The high scores in Figure 8c show vast areas in Paris that are characterized by the temporal pattern of the overall traffic discussed above. However, Figure 8e also highlights concentrations of negative-score regions that are thus affected by the same patterns in the opposite way: these areas experience higher traffic demands during the work hours and low activity periods during evenings and weekends. Comparing the scores against INSEE data about the density of residential buildings, in Figure 8d, and of commercial and industrial buildings, in Figure 8f, reveals significant spatial similarities. To quantify the relationship, we separate scores above and below 0 as well as the maps of residential areas and of commercial or industrial areas. We then overlap the factor 2 high-score map to the residential areas, and the factor 2 low-score map to the commercial or industrial areas, computing in both cases the F1 score that measures the predictive power of factor 2 in identifying each type of area. We find high F1 scores of 0.68 and 0.54 for Paris, and similar values for the other cities, indicating an important connection between the factor and the residential versus commercial nature of the underlying region.

Takeaway. Neighborhoods characterized by residential activities and work dynamics are told apart by how traffic fluctuates between 8 am and 6 pm during working days: residential areas see a reduced demand in those periods, while traffic surges in commercial areas. As the effect is homogeneous across services, total traffic is a good indicator for this factor, which explains why studies based on aggregated demands primarily identify this work-life dichotomy [40], [41].

C. Factor 4, or the digital footprint of budget nightlife

Like the previous factors, also factor 4 describes usage patterns that characterize the vast majority of applications.



(a) Factor 1 scores (b) Income (c) Factor 2 high scores (d) Residential (e) Factor 2 low scores (f) Commercial

Fig. 8: Scores for (a) factor 1 and (c,e) factor 2. Maps of (b) normalized income, (d) residential and (f) commercial areas.

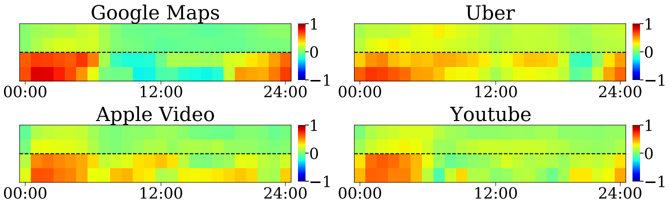
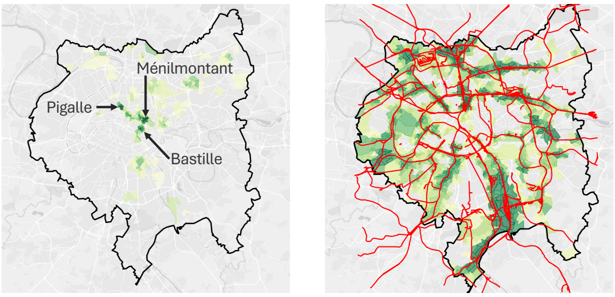


Fig. 9: Factor 4 loadings for selected applications. In each plot, SRCA (resp., traffic) loadings are top (resp., bottom) two rows.



(a) Factor 4 (b) Factor 10

Fig. 10: High-scores IRIS zones for factors 4 and 10 in Paris.

The loadings in Figure 9 offers a glance to the fact that most services show peaks in their demands during the late evening and night hours, from 10 pm through 6 am. Instead, the relative usage of applications does not vary as shown by near-0 SRCA.

The cartography of scores in Figure 10a pinpoints in fact specific hotspots in Paris that collect the nighttime traffic surge outlined by the loadings. These are well-known nightlife areas in the city, such as Pigalle, Bastille or Ménilmontant, which offer a variety of nightclubs, late-hour bars or adult show theaters. In particular, these areas offer nighttime entertainment at a reasonable cost and are popular for their affordable accommodation. This relates them to nightlife in low-income locations that are also highlighted by the scores, although to a lesser extent, such as low-income Bagnolet or Ivry.

Takeaway. Locations characterized by nightlife activities on a budget see a generalized increased consumption of all mobile services during the late evening and night hours.

D. Factor 6, or a mobile service synopsis of student life

Factor 6 is applications-specific with high L_a^2 for Instagram and Twitter/X. Figure 11 shows that (i) the SRCA loadings reflect a higher usage than normal for these services, while (ii) the scores pinpoint specific locations in the French cities that match university campuses (*e.g.*, Domaine Universitaire in Bordeaux, or Campus de Luminy in Marseille). Interestingly, the scores also reveal precise areas in each city that are popular places for university students to hang out (*e.g.*, Victoire in Bordeaux, or Notre Dame du Mont in Marseille).

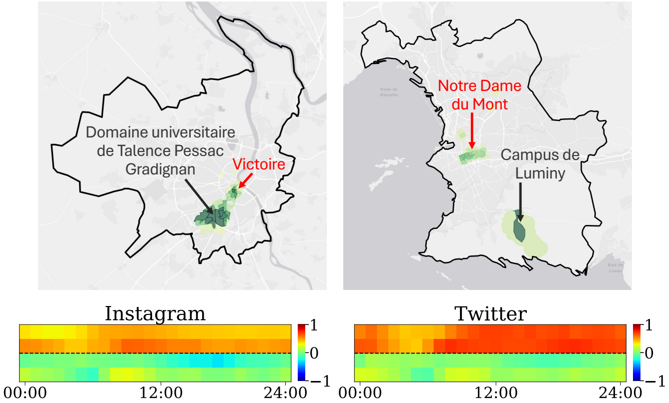


Fig. 11: Factor 6 scores in Bordeaux (left) and Marseille (right), only high scores are shown. Factor 6 loadings for Instagram and Twitter: in each plot, SRCA (resp., traffic) loadings are the top (resp., bottom) two rows.

Takeaway. High relative Instagram and Twitter/X demands are an excellent predictor of university students' presence.

E. Factor 8, or connecting traffic on roads and networks

This factor yields high L_a^2 for mobility-related applications such as Waze and Google Maps, as well as for Apple Siri. The loadings, exemplified in Figure 12, show that such services are used more than usual at all times (*e.g.*, Waze) or during daylight hours (Siri). Scores, in Figure 12, immediately explain the root cause for this factor: overlaying the high-score IRIS zones with the topology of the city road networks exposes a clear match between the factor and the major local roadways.

Takeaway. Highly trafficked roads induce an exceptional consumption of not only of navigation services like Waze or Google Maps but also of personal assistants like Siri. We hypothesize that this may be due to many drivers being more comfortable using speech commands while traveling.

F. Factor 10, or mobile traffic also commutes

Factor 10 concerns a large portion of the applications, which all show peaks of demand at 7-8 am and 6-7 pm, *i.e.*, the home-work commuting times in France, according to the loadings, depicted in Figure 13 for sample services. High-score zones,

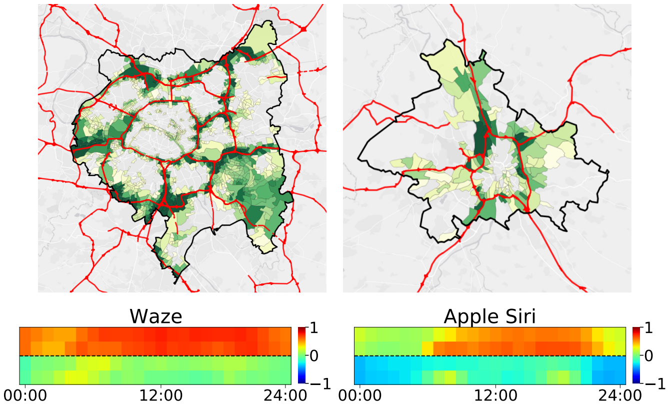


Fig. 12: Factor 8 scores in Paris and Toulouse, only high scores are shown. Factor 8 loadings for Waze and Siri: in each plot, SRCA (resp., traffic) loadings are top (resp., bottom) two rows.

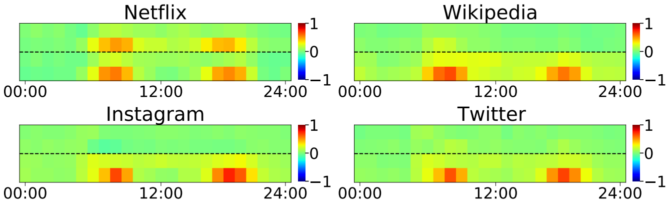


Fig. 13: Factor 10 loadings for selected services. In each plot, SRCA (resp., traffic) loadings are top (resp., bottom) two rows.

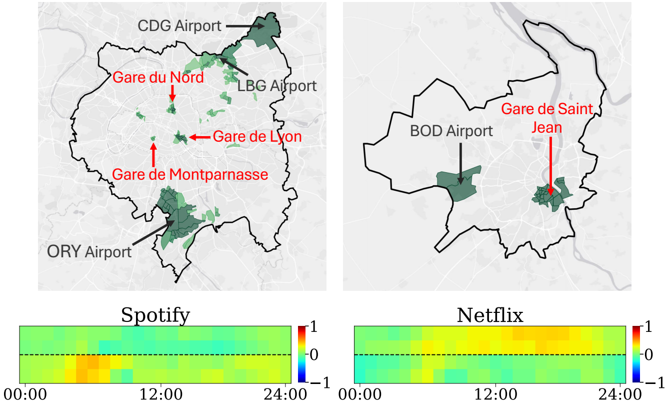


Fig. 14: Factor 14 scores in Paris (left) and Bordeaux (right). Factor 14 loadings for Spotify and Netflix: in each plot, SRCA (resp., traffic) loadings are the top (resp., bottom) two rows.

in Figure 10b, map well to the topology of metropolitan train lines, indicating that the factor describes the traffic induced by medium- and long-range commuters on public transport.

Takeaway. Commuters on metropolitan trains consume a variety of mobile applications, ranging from social media (*e.g.*, Instagram) to knowledge sources (*e.g.*, Wikipedia). Interestingly, video streaming (*e.g.*, Netflix) also shows higher relative usage in SRCA loadings, implying an especially significant burst of usage with respect to the other affected services.

G. Factor 14, or the multiple facets of transportation hubs

Also this factor has high L_a^2 on most services, yet with quite diverse loadings, exemplified in Figure 14. Factor 14 is linked for instance to peaks of audio streaming (*e.g.*, Spotify) traffic early in the morning and higher relative usage of video streaming (*e.g.*, Netflix) during the afternoons. Scores, in Figure 14 help unraveling the factor as they highlight airports or major train stations in all cities.

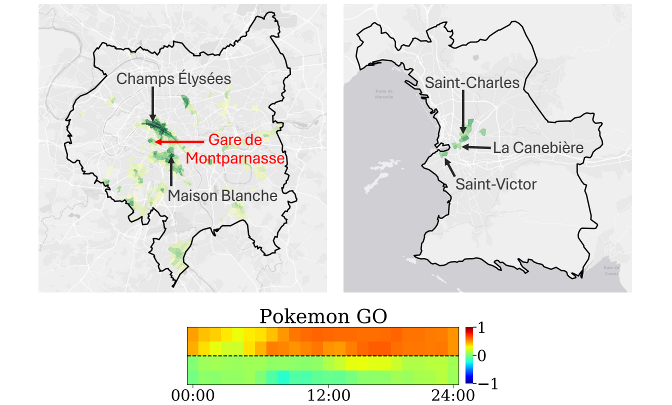


Fig. 15: Factor 15 scores in Paris (left) and Marseille (right). Factor 15 loadings for Pokemon GO: in each plot, SRCA (resp., traffic) loadings are the top (resp., bottom) two rows.

Takeaway. Airports show multi-faceted mobile service demands with unique and diverse application-level dynamics.

H. Factor 15, or the places of augmented-reality gaming

This factor is exclusively associated with Pokémon GO according to L_a^2 values, and indeed pinpoints a higher usage than normal for this service as shown in Figure 15. Scores in the same figure highlight iconic locations and large parks or pedestrian areas in each city, which the developer arguably selects as good areas for a better augmented-reality experience.

Takeaway. Due to the combination of its augmented-reality nature and huge popularity, Pokemon GO is a fairly unique game, to the point that its spatiotemporal dynamics cannot be categorized along those of other mobile services. This result is a strong indicator that applications controlling the user movements can generate original patterns not seen in traditional human activity-driven mobility.

VI. Conclusions

We customized exploratory factor analysis to reveal latent dynamics in service-level mobile traffic and applied such a tool to a real-world dataset collected by a network operator in 10 cities of France. Our results demonstrate the effectiveness of the approach to reveal a number of tangled patterns in space and time that are potentially associated to specific applications. Our work provides both a promising methodology for in-depth traffic characterization and novel insights on how urban fabrics control the demands for mobile services.

Acknowledgment

This work was supported by the ORIGAMI project (GA 101139270), funded by SNS JU and the European Union; CoCo5G (ANR-22-CE25-0016), funded by the French National Research Agency; and CHIST-ERA ECOMOME, funded by the Fonds de Recherche du Québec (FRQNT).

References

[1] S. Mishra *et al.*, "Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in france," in *The World Wide Web Conference*, 2022, pp. 2532–2540.

[2] C. Song *et al.*, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[3] B. C. Csáji *et al.*, "Exploring the mobility of mobile phone users," *Physica A: statistical mechanics and its applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[4] K. S. Kung *et al.*, "Exploring universal patterns in human home-work commuting from mobile phone data," *PloS one*, vol. 9, no. 6, 2014.

[5] T. Louail *et al.*, "Uncovering the spatial structure of mobility networks," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[6] J. E. Steele *et al.*, "Mapping poverty using mobile phone and satellite data," *Journal of The Royal Society Interface*, vol. 14, no. 127, p. 20160690, 2017.

[7] N. Pokhriyal and D. C. Jacques, "Combining disparate data sources for improved poverty prediction and mapping," *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, pp. E9783–E9792, 2017.

[8] E. Moro *et al.*, "Mobility patterns are associated with experienced income segregation in large us cities," *Nat Commun*, vol. 12, 2021.

[9] A. Vazquez Brust *et al.*, "Detecting areas of potential high prevalence of chagas in argentina," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 262–271.

[10] N. Oliver *et al.*, "Mobile phone data for informing public health actions across the covid-19 pandemic life cycle," p. eabc0764, 2020.

[11] A. F. Zanella *et al.*, "Impact of later-stages covid-19 response measures on spatiotemporal mobile service usage," in *IEEE INFOCOM 2022*. IEEE, 2022, pp. 970–979.

[12] D. A. Rohlinger and S. Sobieraj, *The Oxford Handbook of Digital Media Sociology*. Oxford University Press, 09 2022.

[13] K. Park *et al.*, "Technology trends and challenges in sdn and service assurance for end-to-end network slicing," *Computer Networks*, vol. 234, p. 109908, 2023.

[14] U. Paul *et al.*, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM 2021*. IEEE, 2011, pp. 882–890.

[15] F. Li *et al.*, "Who is the king of the hill? traffic analysis over a 4g network," in *IEEE ICC 2018*. IEEE, 2018, pp. 1–6.

[16] P. Parastar *et al.*, "Spotlight on 5g: Performance, device evolution and challenges from a mobile operator perspective," in *IEEE INFOCOM 2023*, 2023, pp. 1–10.

[17] S. Mishra *et al.*, "Characterizing 5g adoption and its impact on network traffic and mobile service consumption," in *IEEE INFOCOM*, 2024.

[18] I. Trestian *et al.*, "Measuring serendipity: connecting people, locations and interests in a mobile 3g network," in *ACM IMC 2009*, 2009, pp. 267–279.

[19] Q. Xu *et al.*, "Identifying diverse usage behaviors of smartphone apps," in *ACM IMC 2011*, 2011, pp. 329–344.

[20] M. Z. Shafiq *et al.*, "Characterizing and modeling internet traffic dynamics of cellular devices," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, pp. 265–276, 2011.

[21] M. Z. Shafiq *et al.*, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *IEEE INFOCOM 2012*. IEEE, 2012, pp. 1341–1349.

[22] R. Keralapura *et al.*, "Profiling users in a 3g network using hourglass co-clustering," in *ACM MobiCom 2010*, 2010, pp. 341–352.

[23] H. Li *et al.*, "Characterizing smartphone usage patterns from millions of android users," in *ACM IMC 2015*, 2015, pp. 459–472.

[24] J. Yang *et al.*, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 95–106, 2015.

[25] P. Fiadino *et al.*, "Vivisecting WhatsApp through Large-Scale Measurements in Mobile Networks," *ACM SIGCOMM Computer Communication Review*, vol. 44(4), Aug. 2014.

[26] P. Fiadino *et al.*, "Online social networks anatomy: On the analysis of facebook and whatsapp in cellular networks," in *2015 IFIP Networking Conference (IFIP Networking)*, 2015, pp. 1–9.

[27] Q. Deng *et al.*, "An Empirical Study of the WeChat Mobile Instant Messaging Service," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 390–395.

[28] V. Soto and E. Frías-Martínez, "Automated land use identification using cell-phone records," in *Proceedings of the 3rd ACM international workshop on MobiArch*, 2011, pp. 17–22.

[29] J. L. Toole *et al.*, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD international workshop on urban computing*, 2012, pp. 1–8.

[30] M. Lenormand *et al.*, "Comparing and modelling land use organization in cities," *Royal Society open science*, vol. 2, no. 12, p. 150449, 2015.

[31] S. Grauwin *et al.*, "Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong," *Computational approaches for urban environments*, pp. 363–387, 2015.

[32] Z. Xing *et al.*, "Inferring land use type in urban area with mobile big data," in *IEEE ICCC 2018*. IEEE, 2018, pp. 1835–1840.

[33] Z. Sun *et al.*, "Deep convolutional autoencoder for urban land use classification using mobile device data," *International Journal of Geographical Information Science*, vol. 36, no. 11, pp. 2138–2168, 2022.

[34] M. De Nadai *et al.*, "The death and life of great italian cities: a mobile phone data perspective," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 413–423.

[35] S. Bakirtzis *et al.*, "Characterizing mobile service demands at indoor cellular networks," in *ACM IMC 2023*, 2023, pp. 645–659.

[36] Z. Tu *et al.*, "Demographics of mobile app usage: long-term analysis of mobile app usage," *CCF Trans. Pervasive Comp. Interact.*, vol. 3, p. 235–252, 2021.

[37] T. Li *et al.*, "Finding spatiotemporal patterns of mobile application usage," *IEEE Trans. Netw. Sci. Eng*, pp. 1–1, 2021.

[38] D. Yu *et al.*, "Smartphone app usage prediction using points of interest," *ACM IMWUT 2018*, vol. 1, no. 4, pp. 1–21, 2018.

[39] H. Wang *et al.*, "Modeling spatio-temporal app usage for a large user population," *ACM IMWUT 2019*, vol. 3, no. 1, pp. 1–23, 2019.

[40] A. Furno *et al.*, "Joint spatial and temporal classification of mobile traffic demands," in *IEEE INFOCOM 2017*, 2017, pp. 1–9.

[41] A. Furno *et al.*, "Spatial and temporal exploratory factor analysis of urban mobile data traffic," *Data Science for Transportation*, vol. 6, no. 1, p. 4, 2024.

[42] C. Marquez *et al.*, "Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage," in *ACM CoNEXT 2017*, 2017, pp. 180–186.

[43] R. Singh *et al.*, "Urban vibes and rural charms: Analysis of geographic diversity in mobile service usage at national scale," in *The World Wide Web Conference*, 2019, pp. 1724–1734.

[44] E. Union. (2016) Eu general data protection regulation (gdpr): Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). [Online]. Available: https://gdpr-info.eu/

[45] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[46] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996.

[47] C. E. Brown, "Coefficient of variation," in *Applied multivariate statistics in geohydrology and related sciences*. Springer, 1998, pp. 155–157.

[48] M. W. Watkins, "Exploratory factor analysis: A guide to best practice," *Journal of black psychology*, vol. 44, no. 3, pp. 219–246, 2018.

[49] B. Balassa, "Trade liberalisation and "revealed" comparative advantage 1," *The manchester school*, vol. 33, no. 2, pp. 99–123, 1965.

[50] K. Y. Hogarty *et al.*, "The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination," *Educational and psychological measurement*, vol. 65, 2005.

[51] R. C. MacCallum *et al.*, "Sample size in factor analysis." *Psychological methods*, vol. 4, no. 1, p. 84, 1999.

[52] H. F. Kaiser and J. Rice, "Little jiffy, mark iv," *Educational and psychological measurement*, vol. 34, no. 1, pp. 111–117, 1974.

[53] I. T. Jolliffe, "Principal component analysis and factor analysis," *Principal component analysis*, pp. 150–166, 2002.

[54] J. C. F. de Winter and D. Dodou, "Common factor analysis versus principal component analysis: A comparison of loadings by means of simulations," *Communications in Statistics - Simulation and Computation*, vol. 45, no. 1, pp. 299–321, 2016.

[55] L. R. Tucker and C. Lewis, "A reliability coefficient for maximum likelihood factor analysis," *Psychometrika*, vol. 38, no. 1, pp. 1–10, 1973.

[56] H. Taherdoost *et al.*, "Exploratory factor analysis; concepts and theory," *Advances in applied and pure mathematics*, vol. 27, pp. 375–382, 2022.

[57] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[58] I. Ucar *et al.*, "News or social media? socio-economic divide of mobile service consumption," *Journal of The Royal Society Interface*, vol. 18, no. 185, p. 20210350, 2021.